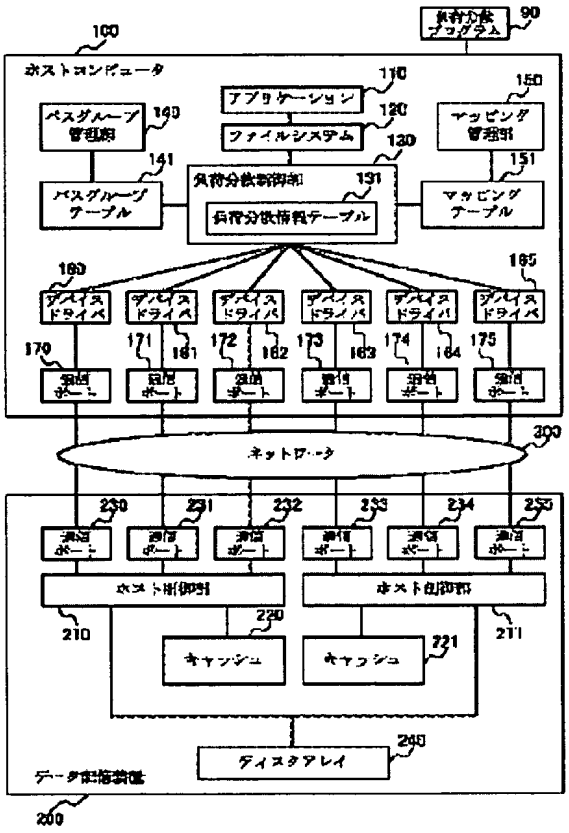
LOAD-SHARING SYSTEM, HOST COMPUTER FOR THE LOAD-SHARING SYSTEM, AND LOAD-SHARING PROGRAM

Patent number:	JP2003099384
Publication date:	2003-04-04
Inventor:	KATSURAJIMA KO
Applicant:	NIPPON ELECTRIC CO
Classification:	
- international:	G06F13/14; G06F3/06; G06F12/08
- european:	
Application number:	JP20010286952 20010920
Priority number(s):	JP20010286952 20010920

Report a data error here

Abstract of **JP2003099384**

PROBLEM TO BE SOLVED: To provide a multi-path load-sharing system in a host computer, which accesses a data storage device with unshared cache method. SOLUTION: In the load-sharing system, connecting a host computer 100 and a data storage device 200 with the unshared cache method via a network 300, a host computer 100 is constituted with a load-sharing control part 130 which groups plural physical paths with the data storage device 200 for each of caches which are to be used by a host control part, for executing prefetch within the data storage device 200 and manage them as a path group, map a sequential readout request into a single path group and properly distributing input output requests into the plural physical paths.



(19)日本国特許庁（ＪＰ）　(12) **公開特許公報**（Ａ）　(11)特許出願公開番号
特開2003－99384
（P2003－99384A）
(43)公開日　平成15年４月４日(2003.4.4)

(51)Int.Cl.[7]	識別記号	ＦＩ		テーマコード*（参考）
Ｇ０６Ｆ　13/14	３１０	Ｃ０６Ｆ　13/14	３１０Ｈ	５Ｂ００５
3/06	３０２	3/06	３０２Ａ	５Ｂ０１４
12/08	５１１	12/08	５１１Ｚ	５Ｂ０６５
	５１５		５１５Ｚ	
	５１９		５１９Ｚ	

審査請求　未請求　請求項の数22　ＯＬ　（全 17 頁）　最終頁に続く

(21)出願番号　特願2001－286952(P2001－286952)

(22)出願日　平成13年９月20日(2001.9.20)

(71)出願人　000004237
日本電気株式会社
東京都港区芝五丁目７番１号

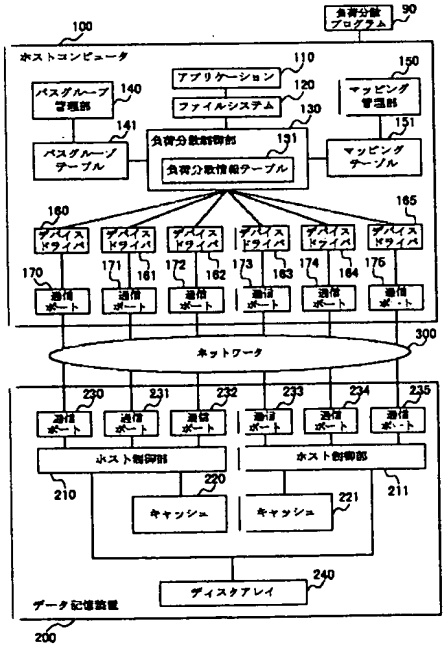
(72)発明者　桂島　航
東京都港区芝五丁目７番１号　日本電気株式会社内

(74)代理人　100093595
弁理士　松本　正夫

Ｆターム(参考)　5B005 JJ13 MM12 NN12
5B014 HA13
5B065 CC08 CE13 ZA08

(54)【発明の名称】　負荷分散システム、負荷分散システムのホストコンピュータ、及び負荷分散プログラム

(57)【要約】　（修正有）
【課題】　非共有キャッシュ方式のデータ記憶装置にアクセスするホストコンピュータにおいてマルチパス負荷分散システムを提供する。
【解決手段】　ホストコンピュータ１００と、非共有キャッシュ方式のデータ記憶装置２００をネットワーク３００を介して接続する負荷分散システムにおいて、ホストコンピュータ１００は、データ記憶装置２００との間における複数の物理的パスを、データ記憶装置２００内においてプリフェッチを実行するホスト制御部が使用するキャッシュ毎にグループ化してパスグループとして管理し、シーケンシャルな読み出し要求を単一のパスグループにマッピングし、入出力要求を複数本の物理的パスに適切に分散する負荷分散制御部１３０を備える。



【特許請求の範囲】

【請求項１】 ホストコンピュータと、非共有キャッシュ方式のデータ記憶装置をネットワークを介して接続する負荷分散システムにおいて、

入出力要求のアドレス情報と、前記データ記憶装置との間における複数の物理的パスとのマッピングを行ない、シーケンシャルな読み出し要求を単一の前記パスにマッピングし、前記マッピングの内容を示すマッピング情報を生成し、

前記マッピング情報を参照して、前記入出力要求を複数本の物理的パスに適切に分散することを特徴とする負荷分散システム。

【請求項２】 ホストコンピュータと、非共有キャッシュ方式のデータ記憶装置をネットワークを介して接続する負荷分散システムにおいて、

前記データ記憶装置との間における複数の物理的パスを、前記データ記憶装置内のキャッシュ毎にグループ化してパスグループとして管理し、前記パスグループの情報を示すパスグループ情報を生成し、

入出力要求のアドレス情報と前記パスグループとのマッピングを行ない、シーケンシャルな読み出し要求を単一の前記パスグループにマッピングし、前記マッピングの内容を示すマッピング情報を生成し、

前記パスグループ情報と前記マッピング情報を参照して、前記入出力要求を複数本の物理的パスに適切に分散することを特徴とする負荷分散システム。

【請求項３】 ホストコンピュータと、非共有キャッシュ方式のデータ記憶装置をネットワークを介して接続する負荷分散システムにおいて、

前記データ記憶装置と前記ホストコンピュータとの間の通信を中継し、前記データ記憶装置のキャッシングを実行するホスト制御装置を備え、

入出力要求のアドレス情報と、前記ホストコンピュータと前記ホスト制御装置との間における複数の物理的パスとのマッピングを行ない、シーケンシャルな読み出し要求を単一の前記パスにマッピングし、前記マッピングの内容を示すマッピング情報を生成し、

前記マッピング情報を参照して、前記入出力要求を複数本の物理的パスに適切に分散することを特徴とする負荷分散システム。

【請求項４】 ホストコンピュータと、非共有キャッシュ方式のデータ記憶装置をネットワークを介して接続する負荷分散システムにおいて、

前記データ記憶装置と前記ホストコンピュータとの間の通信を中継し、前記データ記憶装置のキャッシングを実行するホスト制御装置を備え、

前記ホストコンピュータと前記ホスト制御装置との間における複数の物理的パスを、前記ホスト制御装置内のキャッシュ毎にグループ化してパスグループとして管理し、前記パスグループの情報を示すパスグループ情報を生成し、

入出力要求のアドレス情報と前記パスグループとのマッピングを行ない、シーケンシャルな読み出し要求を単一の前記パスグループにマッピングし、前記マッピングの内容を示すマッピング情報を生成し、

前記マッピング情報を参照して、前記入出力要求を複数本の物理的パスに適切に分散することを特徴とする負荷分散システム。

【請求項５】 前記マッピング情報の生成において、前記入出力要求のアドレス情報として、当該入出力要求の論理ボリューム番号を用いることを特徴とする請求項１から請求項４のいずれか１つに記載の負荷分散システム。

【請求項６】 前記入出力要求の論理アドレス番号を適当な幅毎に区画化し、

前記マッピング情報の生成において、前記入出力要求のアドレス情報として、当該入出力要求の論理アドレス番号が該当する前記区画を用いることを特徴とする請求項１から請求項４のいずれか１つに記載の負荷分散システム。

【請求項７】 前記ネットワークを介して、前記入出力要求の論理アドレス番号の区画化に用いる最適な区画幅の情報を得る手段を備えることを特徴とする請求項６記載の負荷分散システム。

【請求項８】 ホストコンピュータと、非共有キャッシュ方式のデータ記憶装置をネットワークを介して接続する負荷分散システムにおいて、

前記ホストコンピュータが、前記データ記憶装置との間における複数の物理的パスを、前記データ記憶装置内のキャッシュ毎にグループ化してパスグループとして管理するパスグループ情報を記録するパスグループテーブルと、シーケンシャルな読み出し要求を単一の前記パスグループにマッピングした、入出力要求のアドレス情報と前記パスグループとのマッピングを示すマッピング情報を記録するマッピングテーブルを備え、

前記パスグループ情報と前記マッピング情報を参照して、前記入出力要求を複数本の物理的パスに適切に分散することを特徴とする負荷分散システム。

【請求項９】 非共有キャッシュ方式のデータ記憶装置とネットワークを介して接続する負荷分散システムのホストコンピュータにおいて、

入出力要求のアドレス情報と、前記データ記憶装置との間における複数の物理的パスとのマッピングを行ない、シーケンシャルな読み出し要求を単一の前記パスにマッピングし、前記マッピングの内容を示すマッピング情報を生成するマッピング管理手段と、

前記マッピング情報を参照して、前記入出力要求を複数本の物理的パスに適切に分散する負荷分散制御手段を備えることを特徴とするホストコンピュータ。

【請求項１０】 非共有キャッシュ方式のデータ記憶装

置とネットワークを介して接続する負荷分散システムのホストコンピュータにおいて、
前記データ記憶装置との間における複数の物理的パスを、前記データ記憶装置内のキャッシュ毎にグループ化してパスグループとして管理し、前記パスグループの情報を示すパスグループ情報を生成するパスグループ管理手段と、
入出力要求のアドレス情報と前記パスグループとのマッピングを行ない、シーケンシャルな読み出し要求を単一の前記パスグループにマッピングし、前記マッピングの内容を示すマッピング情報を生成するマッピング管理手段と、
前記パスグループ情報と前記マッピング情報を参照して、前記入出力要求を複数本の物理的パスに適切に分散する負荷分散制御手段を備えることを特徴とするホストコンピュータ。
【請求項１１】　前記パスグループ管理手段は、
前記データ記憶装置から、当該データ記憶装置が記憶するデータの構成情報を取得する手段を備えることを特徴とする請求項１０に記載のホストコンピュータ。
【請求項１２】　前記パスグループ管理手段は、
指定された外部の情報機器から、前記データ記憶装置が記憶するデータの構成情報を取得する手段を備えることを特徴とする請求項１０に記載のホストコンピュータ。
【請求項１３】　前記マッピング管理手段は、
前記入出力要求のアドレス情報として、当該入出力要求の論理ボリューム番号を用いてマッピングを行なうことを特徴とする請求項９から請求項１２のいずれか１つに記載のホストコンピュータ。
【請求項１４】　前記マッピング管理手段は、
前記入出力要求の論理アドレス番号を適当な幅毎に区画化し、
前記入出力要求のアドレス情報として、当該入出力要求の論理アドレス番号が該当する前記区画を用いてマッピングを行なうことを特徴とする請求項９から請求項１２のいずれか１つに記載のホストコンピュータ。
【請求項１５】　前記マッピング管理手段は、
前記データ記憶装置から、前記入出力要求の論理アドレス番号の区画化に用いる最適な区画幅の情報を得る手段を備えることを特徴とする請求項１４記載のホストコンピュータ。
【請求項１６】　前記マッピング管理手段は、
指定された外部の情報機器から、前記入出力要求の論理アドレス番号の区画化に用いる最適な区画幅の情報を得る手段を備えることを特徴とする請求項１４記載のホストコンピュータ。
【請求項１７】　前記外部の情報機器を、ネームサーバとすることを特徴とする請求項１２又は請求項１６に記載のホストコンピュータ。
【請求項１８】　前記マッピング管理手段は、
定期的に各前記パスの負荷情報を参照して前記マッピング情報を更新することを特徴とする請求項９から請求項１７のいずれか１つに記載のホストコンピュータ。
【請求項１９】　非共有キャッシュ方式のデータ記憶装置とネットワークを介して接続する負荷分散システムのホストコンピュータにおいて、
前記データ記憶装置との間における複数の物理的パスを、前記データ記憶装置内のキャッシュ毎にグループ化してパスグループとして管理するパスグループ情報を記録するパスグループテーブルと、
シーケンシャルな読み出し要求を単一の前記パスグループにマッピングした、入出力要求のアドレス情報と前記パスグループとのマッピングを示すマッピング情報を記録するマッピングテーブルと、
前記パスグループ情報と前記マッピング情報を参照して、前記入出力要求を複数本の物理的パスに適切に分散する負荷分散制御手段を備えることを特徴とするホストコンピュータ。
【請求項２０】　非共有キャッシュ方式のデータ記憶装置とネットワークを介して接続するホストコンピュータを制御することにより、負荷分散を実行する負荷分散プログラムにおいて、
入出力要求のアドレス情報と、前記データ記憶装置との間における複数の物理的パスとのマッピングを行ない、シーケンシャルな読み出し要求を単一の前記パスにマッピングし、前記マッピングの内容を示すマッピング情報を生成するマッピング管理機能と、
前記マッピング情報を参照して、前記入出力要求を複数本の物理的パスに適切に分散する負荷分散制御機能を備えることを特徴とする負荷分散プログラム。
【請求項２１】　非共有キャッシュ方式のデータ記憶装置とネットワークを介して接続するホストコンピュータを制御することにより、負荷分散を実行する負荷分散プログラムにおいて、
前記データ記憶装置との間における複数の物理的パスを、前記データ記憶装置内のキャッシュ毎にグループ化してパスグループとして管理し、前記パスグループの情報を示すパスグループ情報を生成するパスグループ管理機能と、
入出力要求のアドレス情報と前記パスグループとのマッピングを行ない、シーケンシャルな読み出し要求を単一の前記パスグループにマッピングし、前記マッピングの内容を示すマッピング情報を生成するマッピング管理機能と、
前記パスグループ情報と前記マッピング情報を参照して、前記入出力要求を複数本の物理的パスに適切に分散する負荷分散制御機能を備えることを特徴とする負荷分散プログラム。
【請求項２２】　非共有キャッシュ方式のデータ記憶装置とネットワークを介して接続するホストコンピュータ

を制御することにより、負荷分散を実行する負荷分散プログラムにおいて、
前記ホストコンピュータが、前記データ記憶装置との間における複数の物理的パスを、前記データ記憶装置内のキャッシュ毎にグループ化してパスグループとして管理するパスグループ情報を記録するパスグループテーブルと、シーケンシャルな読み出し要求を単一の前記パスグループにマッピングした、入出力要求のアドレス情報と前記パスグループとのマッピングを示すマッピング情報を記録するマッピングテーブルを備え、
前記パスグループ情報と前記マッピング情報を参照して、前記入出力要求を複数本の物理的パスに適切に分散する負荷分散制御機能を備えることを特徴とする負荷分散プログラム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明はホストコンピュータとデータ記憶装置の間に複数本の物理的アクセスパスが存在する場合のホストコンピュータのマルチパス負荷分散に関し、特に、データ記憶装置が非共有キャッシュ方式の場合の負荷分散システム、負荷分散システムのホストコンピュータ、及び負荷分散プログラムに関する。

【０００２】

【従来の技術】高可用性を求められるシステムの場合、ホストコンピュータとデータ記憶装置のアクセスのために、複数本の物理的パスが設けられる。その理由は、或るパスで障害が発生した際に、異なるパスをその代替パスとして用いるためである。

【０００３】これらの複数本の物理的パスを効率的に利用するために、ホストコンピュータ側でのマルチパス負荷分散が広く行なわれている。データ記憶装置に対するアクセスの負荷を複数本の物理的パスに適切に分散することで、ホストコンピュータとデータ記憶装置の間のスループットを向上させるのが狙いである。

【０００４】マルチパス負荷分散の制御を行なう手段は、ファイルシステムレイヤとデバイスドライバレイヤの間に位置しており、ホストコンピュータのファイルシステムからデータ記憶装置に対するＩ／Ｏ要求（入出力要求）を受け取ると、負荷分散アルゴリズムに基づき使用するパスを決定し、適切なデバイスドライバへＩ／Ｏ要求を渡す仕組みになっている。

【０００５】次に、図面を参照して従来の負荷分散システムの動作を説明する。図１２は、従来の負荷分散システムの構成を示すブロック図である。

【０００６】ホストコンピュータ１００ｅ内のファイルシステム１２０は、アプリケーション１１０からＩ／Ｏ要求を受け取ると、それらをブロック単位のＩ／Ｏ要求に変換し、負荷分散制御部１３０ｅに送付する。負荷分散制御部１３０ｅは、ファイルシステム１２０からＩ／Ｏ要求を受け取ると、負荷分散アルゴリズムに基づき使用するパスを決定する。最後に、デバイスドライバ１６０～１６５の中で適切なものに対し、Ｉ／Ｏ要求を送付する。

【０００７】負荷分散のアルゴリズムとしては、発生するＩ／Ｏ要求を、複数本のパスに対しラウンドロビンにて送付する方法、ペンディングされているＩ／Ｏ要求の数が最も少ないパスに送付する方法、ペンディングされているＩ／Ｏ要求のブロック数の総和が最も少ないパスに送付する方法等が使われている（ここでペンディングされているＩ／Ｏ要求とは、確認応答が返ってきていないＩ／Ｏ要求のことを指している）。

【０００８】また、従来技術の一例として、特開平０７－３３４４４９号公報に開示された技術が挙げられる。この方式は、物理的パスを識別する入出力経路識別部と、各物理的パスの負荷を測定する負荷測定部と、負荷情報を蓄積する負荷統計情報蓄積部と、蓄積された負荷情報に基づきパスを決定する入出力経路決定部を備える。

【０００９】入出力経路決定部は、周辺装置に対する入出力要求時に指定される閾値と、前記負荷統計情報蓄積部によって蓄積された前記統計情報によって示される入出力制御装置の負荷状況とを比較した結果に基づいて、入出力経路の変更が必要か否かを判別する情報提供部と、前記情報提供部によって、負荷状況が閾値を越えて入出力経路の変更が必要と判別された場合に、複数の入出力経路から低負荷の経路を選択する低負荷経路選択部とを具備し、前記低負荷経路選択部によって選択された入出力経路を経由して入出力要求が行なわれることを特徴とする。

【００１０】

【発明が解決しようとする課題】しかし、上述した従来の負荷分散システムでは、非共有キャッシュ方式のデータ記憶装置に対してマルチパス負荷分散を用いると、データ記憶装置のプリフェッチ性能が低下するという問題点があった。以下、この問題点について詳細に説明する。

【００１１】高機能なデータ記憶装置は、ホストコンピュータからのＩ／Ｏ要求をＲＡＩＤ（Redundant Array of Inexpensive Disks）によるディスクアレイで実行する前に、プロセッサ等を用いて処理を行なうことでいくつかの機能を実現している。以下、この処理を行なう部分をホスト制御部と呼ぶことにする。

【００１２】ホスト制御部の機能としては、ホストコンピュータ側に提示するボリュームを柔軟に構成することを可能にする仮想ボリューム機能、スループットやレスポンスタイム向上のためにＲＡＭ（Random Access Memory）等を使って行なわれるキャッシング機能、等が挙げられる。

【００１３】キャッシングは、ホストコンピュータに対して磁気ディスクへのアクセス時間を隠蔽することでス

ループットやレスポンスタイムを向上させる機能である。すなわち、リード要求はキャッシュヒットすれば磁気ディスクにアクセスする必要が無くなるし、ライト要求はとりあえずキャッシングしてその時点でホストコンピュータへ確認応答を返すようにすれば、ホストコンピュータ側からは磁気ディスクへのアクセスはまるで起こっていないかの様に見える。

【００１４】ホスト制御部は、並列に複数個備えられることが多い。その理由は２つある。

【００１５】１つ目は高可用性である。すなわち、或るホスト制御部で障害が発生した際に、異なるホスト制御部をその代替として用いるためである。

【００１６】２つ目は、ボトルネック防止である。すなわち、ホスト制御部の処理能力がシステムの他の能力と比べて低くならないように、同様の処理が行えるホスト制御部を並列に複数配置して処理能力の向上を図るわけである。

【００１７】このように複数のホスト制御部を持つデータ記憶装置の場合、キャッシュの構成方式によって共有キャッシュ方式と非共有キャッシュ方式の２つに分類できる。図１３は、従来の共有キャッシュ方式のデータ記憶装置２００ｆと、非共有キャッシュ方式のデータ記憶装置２００ｇの構成を示すブロック図である。

【００１８】全てのホスト制御部から単一のキャッシュ、単一のキャッシュ制御情報を参照する方式が共有キャッシュ方式であり、いくつかのホスト制御部毎に異なるキャッシュ、異なるキャッシュ制御情報を参照する方式が非共有キャッシュ方式である。

【００１９】ここでキャッシュ制御情報とは、キャッシュしたデータのアドレス情報やキャッシングの順序関係を整理して記した情報のことであり、どのデータをキャッシングするか、又はどのデータを掃き出すかを決定するための情報として使われる。

【００２０】共有キャッシュ方式では、同じアドレス領域のデータを二重にキャッシュすることがないためにキャッシュ容量を有効に使うことができ、またキャッシュは１つだけ用意すればよいから容量を比較的大きくすることができるという利点がある。結果として、キャッシュヒット率が非共有キャッシュ方式に比べて高くなる。欠点は、複数のプロセッサによる競合が発生するため遅延が若干発生することと、複雑な制御を行なうためコストが高くなりがちなことである。

【００２１】一方非共有キャッシュ方式は、複数のプロセッサによる競合が発生しにくい、構造が簡単になりコストを低くできる等の利点がある反面、キャッシュの一貫性の維持という問題がある。

【００２２】この問題の対応策としては、キャッシュ間でダーティデータのみコピーを行なう方法が一般的である。ここでダーティデータとは、後に磁気ディスクにライトされる目的でキャッシングされているデータのことを指している。どこかのキャッシュにデータが書き込まれダーティデータとなった場合、他のキャッシュにコピーを行なうのである。この方法によって、キャッシュの一貫性を保つことができる。

【００２３】また、非共有キャッシュ方式では、キャッシュ制御情報がキャッシュ毎に別々に扱われているために、プリフェッチ等の機能がキャッシュを共有するホスト制御部群を単位として働く。

【００２４】プリフェッチ機能とは、ホストコンピュータからのＩ／Ｏ要求の中からシーケンシャルなリードＩ／Ｏ（読み出し要求）を選り分け、それらに関しては先回りして磁気ディスクからデータを読み出してキャッシュに蓄えておくという機能である。具体的には、近接するアドレスのリードＩ／Ｏが或る程度連続してデータ記憶装置に到着した場合、キャッシュ制御情報からそれらのＩ／ＯがシーケンシャルなリードＩ／Ｏであると判断し、以降のアドレスのデータを適当な分だけ読み出してキャッシュに蓄える。

【００２５】例えば、ホストコンピュータがシーケンシャルなリードＩ／Ｏをデータ記憶装置に送った場合、ホスト制御部は初めの３、４個程度の近接するアドレスのリードＩ／Ｏを受け取った時点でそれ以降のアドレスに対するリードＩ／Ｏを磁気ディスクに送付する。順次ホストコンピュータから到着するリードＩ／Ｏは、キャッシュヒットしレスポンスタイムが大幅に削減されるという仕組みである。

【００２６】本発明が解決する課題は、非共有キャッシュ方式のデータ記憶装置に対してマルチパス負荷分散を用いると、データ記憶装置のプリフェッチ性能が著しく低下するという問題である。

【００２７】例えば、マルチパス負荷分散の負荷分散アルゴリズムにラウンドロビンを用いている場合、ホストコンピュータで発生したシーケンシャルなリードＩ／Ｏは各パスに対して順番に振り分けられるから、複数のキャッシュ制御情報にそれぞれ断続的なアドレスへのリードＩ／Ｏが記録されることになる。この場合、性能低下を招く要因が２点存在する。

【００２８】１点目は、キャッシュ制御情報に連続的なアドレスのリードＩ／Ｏが記録されにくいために、シーケンシャルなリードＩ／Ｏを選り分ける精度が低下するということである。この点は、シーケンシャルなリードＩ／Ｏのキャッシュヒット率を低下させることになる。

【００２９】２点目は、複数のキャッシュ制御情報に共に断続的なアドレスのリードＩ／Ｏが記録されるため、複数のホスト制御部にてシーケンシャルなリードＩ／Ｏだと判断してしまい、同一のアドレス領域へのプリフェッチを多重に行なってしまう可能性があることである。この点は、そのアドレス領域に属する磁気ディスクの混雑を招くことになる。すなわち、シーケンシャルなリードＩ／Ｏのレスポンスタイムを大幅に悪化させてしま

う。

【００３０】上記した例は、負荷分散アルゴリズムに、ラウンドロビン以外の特開平０７－３３４４４９号公報に開示されたような方法を用いた場合でも同様である。これは、非共有キャッシュ方式のデータ記憶装置に対してマルチパス負荷分散を用いると、複数のキャッシュ制御情報にそれぞれ断続的なアドレスへのリードＩ／Ｏが記録されることになるからである。

【００３１】本発明の目的は、上記従来技術の欠点を解決し、非共有キャッシュ方式のデータ記憶装置にアクセスするホストコンピュータが、データ記憶装置のホスト制御部によって行なわれるプリフェッチの性能を低下させることなくマルチパス負荷分散を行なうことができる負荷分散システム、負荷分散システムのホストコンピュータ、及び負荷分散プログラムを提供することにある。

【００３２】特に、ホスト制御部がシーケンシャルなリードＩ／Ｏを選り分ける精度が低下する問題、複数のホスト制御部から同一のアドレス領域へのプリフェッチを多重に行なってしまう問題を発生させずにマルチパス負荷分散を行なうことができる負荷分散システム、負荷分散システムのホストコンピュータ、及び負荷分散プログラムを提供することにある。

【００３３】

【課題を解決するための手段】上記目的を達成するため本発明の負荷分散システムは、ホストコンピュータと、非共有キャッシュ方式のデータ記憶装置をネットワークを介して接続する負荷分散システムにおいて、入出力要求のアドレス情報と、前記データ記憶装置との間における複数の物理的パスとのマッピングを行ない、シーケンシャルな読み出し要求を単一の前記パスにマッピングし、前記マッピングの内容を示すマッピング情報を生成し、前記マッピング情報を参照して、前記入出力要求を複数本の物理的パスに適切に分散することを特徴とする。

【００３４】請求項２の本発明の負荷分散システムは、ホストコンピュータと、非共有キャッシュ方式のデータ記憶装置をネットワークを介して接続する負荷分散システムにおいて、前記データ記憶装置との間における複数の物理的パスを、前記データ記憶装置内のキャッシュ毎にグループ化してパスグループとして管理し、前記パスグループの情報を示すパスグループ情報を生成し、入出力要求のアドレス情報と前記パスグループとのマッピングを行ない、シーケンシャルな読み出し要求を単一の前記パスグループにマッピングし、前記マッピングの内容を示すマッピング情報を生成し、前記パスグループ情報と前記マッピング情報を参照して、前記入出力要求を複数本の物理的パスに適切に分散することを特徴とする。

【００３５】請求項３の本発明の負荷分散システムは、ホストコンピュータと、非共有キャッシュ方式のデータ記憶装置をネットワークを介して接続する負荷分散システムにおいて、前記データ記憶装置と前記ホストコンピュータとの間の通信を中継し、前記データ記憶装置のキャッシングを実行するホスト制御装置を備え、入出力要求のアドレス情報と、前記ホストコンピュータと前記ホスト制御装置との間における複数の物理的パスとのマッピングを行ない、シーケンシャルな読み出し要求を単一の前記パスにマッピングし、前記マッピングの内容を示すマッピング情報を生成し、前記マッピング情報を参照して、前記入出力要求を複数本の物理的パスに適切に分散することを特徴とする。

【００３６】請求項４の本発明の負荷分散システムは、ホストコンピュータと、非共有キャッシュ方式のデータ記憶装置をネットワークを介して接続する負荷分散システムにおいて、前記データ記憶装置と前記ホストコンピュータとの間の通信を中継し、前記データ記憶装置のキャッシングを実行するホスト制御装置を備え、前記ホストコンピュータと前記ホスト制御装置との間における複数の物理的パスを、前記ホスト制御装置内のキャッシュ毎にグループ化してパスグループとして管理し、前記パスグループの情報を示すパスグループ情報を生成し、入出力要求のアドレス情報と前記パスグループとのマッピングを行ない、シーケンシャルな読み出し要求を単一の前記パスグループにマッピングし、前記マッピングの内容を示すマッピング情報を生成し、前記マッピング情報を参照して、前記入出力要求を複数本の物理的パスに適切に分散することを特徴とする。

【００３７】請求項５の本発明の負荷分散システムは、前記マッピング情報の生成において、前記入出力要求のアドレス情報として、当該入出力要求の論理ボリューム番号を用いることを特徴とする。

【００３８】請求項６の本発明の負荷分散システムは、前記入出力要求の論理アドレス番号を適当な幅毎に区画化し、前記マッピング情報の生成において、前記入出力要求のアドレス情報として、当該入出力要求の論理アドレス番号が該当する前記区画を用いることを特徴とする。

【００３９】請求項７の本発明の負荷分散システムは、前記ネットワークを介して、前記入出力要求の論理アドレス番号の区画化に用いる最適な区画幅の情報を得る手段を備えることを特徴とする。

【００４０】請求項８の本発明の負荷分散システムは、ホストコンピュータと、非共有キャッシュ方式のデータ記憶装置をネットワークを介して接続する負荷分散システムにおいて、前記ホストコンピュータが、前記データ記憶装置との間における複数の物理的パスを、前記データ記憶装置内のキャッシュ毎にグループ化してパスグループとして管理するパスグループ情報を記録するパスグループテーブルと、シーケンシャルな読み出し要求を単一の前記パスグループにマッピングした、入出力要求のアドレス情報と前記パスグループとのマッピングを示す

マッピング情報を記録するマッピングテーブルを備え、前記パスグループ情報と前記マッピング情報を参照して、前記入出力要求を複数本の物理的パスに適切に分散することを特徴とする。

【００４１】請求項９の本発明のホストコンピュータは、非共有キャッシュ方式のデータ記憶装置とネットワークを介して接続する負荷分散システムのホストコンピュータにおいて、入出力要求のアドレス情報と、前記データ記憶装置との間における複数の物理的パスとのマッピングを行ない、シーケンシャルな読み出し要求を単一の前記パスにマッピングし、前記マッピングの内容を示すマッピング情報を生成するマッピング管理手段と、前記マッピング情報を参照して、前記入出力要求を複数本の物理的パスに適切に分散する負荷分散制御手段を備えることを特徴とする。

【００４２】請求項１０の本発明のホストコンピュータは、非共有キャッシュ方式のデータ記憶装置とネットワークを介して接続する負荷分散システムのホストコンピュータにおいて、前記データ記憶装置との間における複数の物理的パスを、前記データ記憶装置内のキャッシュ毎にグループ化してパスグループとして管理し、前記パスグループの情報を示すパスグループ情報を生成するパスグループ管理手段と、入出力要求のアドレス情報と前記パスグループとのマッピングを行ない、シーケンシャルな読み出し要求を単一の前記パスグループにマッピングし、前記マッピングの内容を示すマッピング情報を生成するマッピング管理手段と、前記パスグループ情報と前記マッピング情報を参照して、前記入出力要求を複数本の物理的パスに適切に分散する負荷分散制御手段を備えることを特徴とする。

【００４３】請求項１１の本発明のホストコンピュータは、前記パスグループ管理手段は、前記データ記憶装置から、当該データ記憶装置が記憶するデータの構成情報を取得する手段を備えることを特徴とする。

【００４４】請求項１２の本発明のホストコンピュータは、前記パスグループ管理手段は、指定された外部の情報機器から、前記データ記憶装置が記憶するデータの構成情報を取得する手段を備えることを特徴とする。

【００４５】請求項１３の本発明のホストコンピュータは、前記マッピング管理手段は、前記入出力要求のアドレス情報として、当該入出力要求の論理ボリューム番号を用いてマッピングを行なうことを特徴とする。

【００４６】請求項１４の本発明のホストコンピュータは、前記マッピング管理手段は、前記入出力要求の論理アドレス番号を適当な幅毎に区画化し、前記入出力要求のアドレス情報として、当該入出力要求の論理アドレス番号が該当する前記区画を用いてマッピングを行なうことを特徴とする。

【００４７】請求項１５の本発明のホストコンピュータは、前記マッピング管理手段は、前記データ記憶装置から、前記入出力要求の論理アドレス番号の区画化に用いる最適な区画幅の情報を得る手段を備えることを特徴とする。

【００４８】請求項１６の本発明のホストコンピュータは、前記マッピング管理手段は、指定された外部の情報機器から、前記入出力要求の論理アドレス番号の区画化に用いる最適な区画幅の情報を得る手段を備えることを特徴とする。

【００４９】請求項１７の本発明のホストコンピュータは、前記外部の情報機器を、ネームサーバとすることを特徴とする。

【００５０】請求項１８の本発明のホストコンピュータは、前記マッピング管理手段は、定期的に各前記パスの負荷情報を参照して前記マッピング情報を更新することを特徴とする。

【００５１】請求項１９の本発明のホストコンピュータは、非共有キャッシュ方式のデータ記憶装置とネットワークを介して接続する負荷分散システムのホストコンピュータにおいて、前記データ記憶装置との間における複数の物理的パスを、前記データ記憶装置内のキャッシュ毎にグループ化してパスグループとして管理するパスグループ情報を記録するパスグループテーブルと、シーケンシャルな読み出し要求を単一の前記パスグループにマッピングした、入出力要求のアドレス情報と前記パスグループとのマッピングを示すマッピング情報を記録するマッピングテーブルと、前記パスグループ情報と前記マッピング情報を参照して、前記入出力要求を複数本の物理的パスに適切に分散する負荷分散制御手段を備えることを特徴とする。

【００５２】請求項２０の本発明の負荷分散プログラムは、非共有キャッシュ方式のデータ記憶装置とネットワークを介して接続するホストコンピュータを制御することにより、負荷分散を実行する負荷分散プログラムにおいて、入出力要求のアドレス情報と、前記データ記憶装置との間における複数の物理的パスとのマッピングを行ない、シーケンシャルな読み出し要求を単一の前記パスにマッピングし、前記マッピングの内容を示すマッピング情報を生成するマッピング管理機能と、前記マッピング情報を参照して、前記入出力要求を複数本の物理的パスに適切に分散する負荷分散制御機能を備えることを特徴とする。

【００５３】請求項２１の本発明の負荷分散プログラムは、非共有キャッシュ方式のデータ記憶装置とネットワークを介して接続するホストコンピュータを制御することにより、負荷分散を実行する負荷分散プログラムにおいて、前記データ記憶装置との間における複数の物理的パスを、前記データ記憶装置内のキャッシュ毎にグループ化してパスグループとして管理し、前記パスグループの情報を示すパスグループ情報を生成するパスグループ管理機能と、入出力要求のアドレス情報と前記パスグル

ープとのマッピングを行ない、シーケンシャルな読み出し要求を単一の前記パスグループにマッピングし、前記マッピングの内容を示すマッピング情報を生成するマッピング管理機能と、前記パスグループ情報と前記マッピング情報を参照して、前記入出力要求を複数本の物理的パスに適切に分散する負荷分散制御機能を備えることを特徴とする。

【００５４】請求項２２の本発明の負荷分散プログラムは、非共有キャッシュ方式のデータ記憶装置とネットワークを介して接続するホストコンピュータを制御することにより、負荷分散を実行する負荷分散プログラムにおいて、前記ホストコンピュータが、前記データ記憶装置との間における複数の物理的パスを、前記データ記憶装置内のキャッシュ毎にグループ化してパスグループとして管理するパスグループ情報を記録するパスグループテーブルと、シーケンシャルな読み出し要求を単一の前記パスグループにマッピングした、入出力要求のアドレス情報と前記パスグループとのマッピングを示すマッピング情報を記録するマッピングテーブルを備え、前記パスグループ情報と前記マッピング情報を参照して、前記入出力要求を複数本の物理的パスに適切に分散する負荷分散制御機能を備えることを特徴とする。

【００５５】本発明では、パスグループ管理部とマッピング管理部とにより、シーケンシャルな読み出し要求を単一のキャッシュに送付する制限を加える。このような制限を加えることで、シーケンシャルな読み出し要求は、キャッシュを共有しているホスト制御部群によって処理されるようになる。すなわち、その上で負荷分散を行なうことにより、ホスト制御部がシーケンシャルな読み出し要求を選り分ける精度が低下する問題、複数のホスト制御部から同一のアドレス領域へのプリフェッチを多重に行なってしまう問題を解消することができる。

【００５６】また、本発明の第２の実施の形態では、パスグループの情報と、入出力要求のアドレス情報とパスグループとのマッピング情報を、管理者が前もってパスグループテーブル、マッピングテーブルとしてホストコンピュータに与えておくことにより、負荷分散制御部がこれらのテーブルを参照し、負荷分散アルゴリズムに基づき使用するパスを決定する。

【００５７】

【発明の実施の形態】以下、本発明の実施の形態について図面を参照して詳細に説明する。

【００５８】図１は、本発明の第１の実施の形態による負荷分散システムの構成を示すブロック図である。図１を参照すると、本発明の第１の実施の形態は、プログラム制御により動作するホストコンピュータ１００と、非共有キャッシュ方式のデータ記憶装置２００、ネットワーク３００から構成される。

【００５９】ホストコンピュータ１００は、アプリケーション１１０と、ファイルシステム１２０と、負荷分散制御部１３０と、負荷分散情報テーブル１３１と、パスグループ管理部１４０と、パスグループテーブル１４１と、マッピング管理部１５０と、マッピングテーブル１５１と、デバイスドライバ１６０～１６５と、通信ポート１７０～１７５を含む。

【００６０】データ記憶装置２００は、ホスト制御部２１０、２１１と、キャッシュ２２０、２２１と、通信ポート２３０～２３５と、ディスクアレイ２４０を含む。

【００６１】なお、本形態ではアプリケーション１１０、ファイルシステム１２０は、簡便のため１つのみ示したが、複数のアプリケーション、複数のファイルシステムが単一のホストコンピュータ上で動作している場合も考えられる。

【００６２】またデバイスドライバ１６０～１６５、通信ポート１７０～１７５、２３０～２３５（ホストコンピュータ側、データ記憶装置側双方）は、簡便のため６つずつ示したが、それぞれ２～５０程度備えている場合も考えられる。

【００６３】データ記憶装置２００は、簡便のため１つのみ示したが、複数のデータ記憶装置と接続されている場合も考えられる。

【００６４】ホスト制御部２１０、２１１は、簡便のため２つのみ示したが、２～２０程度備えている場合も考えられる。また、各ホスト制御部２１０、２１１は３つずつ通信ポートを持っているが、１～１０程度備えている場合も考えられるし、ホスト制御部間２１０、２１１で異なる数の通信ポートを持つ場合や、異なる性能の通信ポートが混在する場合も考えられる。また、この図１では、１つのホスト制御部が１つのキャッシュを持つ構成になっているが、複数のホスト制御部が１つのキャッシュを共有する場合も考えられる。

【００６５】また、ネットワーク３００は、いろいろな形態が考慮できる。具体的な例としてはFiber Channel Arbitrated Loop、Fiber Channel Fabric、Ethernet、Parallel SCSI（Small Computer System Interface）、ＡＴＭ（Asynchronous Transfer Mode）、ＦＤＤＩ（Fiber Distributed Data Interface）等のデータリンクから構成されることが考えられる。また、途中経路にリピーター、ブリッジ、ルーター、ゲートウェイ等の相互接続機器が設けられていることも考えられる。

【００６６】これらの各部は、それぞれ概略つぎのように動作する。

【００６７】ファイルシステム１２０は、アプリケーション１１０からＩ／Ｏ要求（入出力要求）を受け取ると、それらをブロック単位のＩ／Ｏ要求に変換し、負荷分散制御部１３０に送付する。各Ｉ／Ｏ要求には、ＬＵＮ、ＬＢＡ等のアドレス情報が付加される。

【００６８】負荷分散制御部１３０は、ファイルシステム１２０からＩ／Ｏ要求を受け取ると、まずＩ／Ｏ要求のアドレス情報とマッピングテーブル１５１を参照して

使用するパスグループを決定する。

【００６９】そして、パスグループテーブル１４１を参照してパスグループに属するパスを確認し、その中から負荷分散アルゴリズムに基づき使用するパスを決定する。最後に、デバイスドライバ１６０～１６５の中の適切なものに対して、Ｉ／Ｏ要求を送付する。

【００７０】パスグループ管理部１４０は、まず各データ記憶装置に対するパスを列挙し、そして各パスが接続されているホスト制御部がデータ記憶装置のどのキャッシュを使用しているかを確認する。そして、使用しているキャッシュ毎にパスをグループ化し、その情報をパスグループテーブル１４１に記述する。

【００７１】マッピング管理部１５０は、まずファイルシステム１２０が担当するＩ／Ｏ要求のアドレス情報の範囲を特定し、アドレス情報とパスグループのマッピングを行なう。そして、その情報をマッピングテーブル１５１に記述する。

【００７２】データ記憶装置２００は、複数のキャッシュ２２０、２２１を備えており、非共有キャッシュ方式を採用している。ダーティデータに関しては、分散されたキャッシュ２２０、２２１間でコピーを行なうことで、キャッシュ間でのキャッシュの一貫性を維持している。それ以外のキャッシングデータに関してはコピーを行なわないので、キャッシングデータの内容はキャッシュ間によって異なる。

【００７３】ホスト制御部２１０、２１１は、通信ポート２３０～２３５からＩ／Ｏ要求を受け取ると、そのアドレス情報を参照して、ディスクアレイ２４０に対してＩ／Ｏ要求を送付する。またキャッシュ２２０、２２１を用いてリードキャッシング、ライトキャッシング、プリフェッチを行なう。

【００７４】ディスクアレイ２４０は、複数の磁気ディスク装置を備え、ホスト制御部２１０、２１１からＩ／Ｏ要求を受け取ると、そのアドレス情報を参照して適切な磁気ディスクへＩ／Ｏ要求を送付する。また、内部に複数のＲＡＩＤコントローラを備え、ＲＡＩＤ装置として動作することも考えられる。

【００７５】次に、図２～図４のフローチャートを参照して本実施の形態の全体の動作について詳細に説明する。

【００７６】図２は、本実施の形態の負荷分散制御部１３０の動作を説明するためのフローチャートである。負荷分散制御部１３０は、まずホストコンピュータのブート時にスクリプト等により立ち上げられる。そして、ファイルシステム１２０よりＩ／Ｏ要求が来るまでは待機している（ステップ２０１）。

【００７７】ファイルシステム１２０よりＩ／Ｏ要求が来ると、そのＩ／Ｏ要求のアドレス情報とマッピングテーブル１５１を参照して使用するパスグループを選択する（ステップ２０２）。マッピングテーブル１５１には、Ｉ／Ｏ要求のアドレス情報であるＬＵＮやＬＢＡからパスグループが一意に選択できるようなマッピング情報が記述されている。

【００７８】次に、パスグループテーブル１４１、負荷分散情報テーブル１３１を参照して使用するパスを選択する（ステップ２０３）。パスグループテーブル１４１には、各パスグループにはどのパスが属するかが記述されており、これを参照することでパスグループに属するパスを確認できる。

【００７９】負荷分散情報テーブル１３１には、各パスの利用状況と、各パスグループにおいてどの負荷分散アルゴリズムを用いるかということが記述されており、パスグループテーブル１４１の情報と組み合わせて使用するパスを選択することができる。

【００８０】ここでいうパスの利用状況とは、各パスグループの中で直前にＩ／Ｏ要求を送付したパス、各パスにてペンディングされているＩ／Ｏ要求の数、各パスにてペンディングされているＩ／Ｏ要求のブロック数の総和等を指す。

【００８１】また、ここでいうパスグループの負荷分散アルゴリズムとは、ラウンドロビンにて送付する方法、ペンディングされているＩ／Ｏ要求の数が最も少ないパスに送付する方法、ペンディングされているＩ／Ｏ要求のブロック数の総和が最も少ないパスに送付する方法等を指す。

【００８２】例えば、負荷分散にラウンドロビン方式を用いている場合の具体的な動作手順を以下に示す。まず、ステップ２０２で選択したパスグループにて用いられる負荷分散アルゴリズムを、負荷分散情報テーブル１３１を参照して調べる（この例ではラウンドロビン方式だという結果が出る）。次に、パスグループテーブル１４１を参照して選択したパスグループに属するパスを調べる。最後に、負荷分散情報テーブル１３１内のパスの利用状況を参照し使用するパスが決定する（この例では直前にＩ／Ｏ要求を送付したパスの情報を参照することで、次のパスを決定できる）。

【００８３】この手順は、異なる負荷分散アルゴリズムを用いている場合でもほぼ同様である。ただ異なる利用状況に関する情報を参照するだけである。

【００８４】使用するパスが決定されると、適切なデバイスドライバへＩ／Ｏ要求を送付する（ステップ２０４）。最後に、新しくＩ／Ｏ要求を送付したことで変化したパスの利用状況を、負荷分散情報テーブル１３１に記述する（ステップ２０５）。この後は、再びファイルシステム１２０よりＩ／Ｏ要求が来るまでは、待機することになる。

【００８５】図３は、本実施の形態のパスグループ管理部１４０の動作を説明するためのフローチャートである。パスグループ管理部１４０は、まずホストコンピュータのブート時にスクリプト等により立ち上げられる。

そして、各データ記憶装置に対するパスを列挙する（ステップ３０１）。

【００８６】パスを列挙する分解能は、エンドツーエンドの通信ポートの組合せ、すなわちホストコンピュータ側の各通信ポートと、そこからアクセス可能なデータ記憶装置側の通信ポートの組合せになる。例えば、ホストコンピュータ側の６個の通信ポートがそれぞれ２個のデータ記憶装置側の通信ポートにアクセス可能ならば、全部で１２個のパスが列挙される。

【００８７】アクセス可能であるという意味は、物理的に可能であるという意味だけでなく、ゾーニング等のアクセスコントロールによる制約も受けていないという意味も含んでいる。つまり、このステップ３０１では、アクセス可能なエンドツーエンドのパスを全て列挙するということである。

【００８８】パスが列挙されたなら、各パスが接続されているホスト制御部がデータ記憶装置のどのキャッシュを使用しているかを確認する（ステップ３０２）。このステップ３０２では、まず各パスがデータ記憶装置側のどの通信ポートを用いているか特定する。そして、その通信ポートを使っているホスト制御部がどのキャッシュを使っているか調べる。

【００８９】例えば、本実施の形態では、通信ポート２３０～２３２を用いている場合はキャッシュ２２０を使用し、通信ポート２３３～２３５を用いている場合はキャッシュ２２１を使用しているといったように判別する。

【００９０】ステップ３０２では、データ記憶装置の構成情報、すなわち、通信ポートと使用しているキャッシュの対応関係の情報をどうやって知るかで２つの方法に分かれる。つまり、データ記憶装置に教えてもらう方法と、ネームサーバ等の第三者に教えてもらう方法とがある。

【００９１】１つ目の方法は、例えば各パスがデータ記憶装置のどの通信ポートを用いているか特定する際に、ホストコンピュータとデータ記憶装置間で通信が発生するから、その際にデータ記憶装置側からホストコンピュータへデータ記憶装置の構成情報を渡すというものである。

【００９２】２つ目は、アクセスしたいデータ記憶装置側の通信ポートを特定する際にネームサーバに問い合せる場合、ネームサーバに対応情報を与えておくことで、ホストコンピュータは、通信ポートと使用するキャッシュとの対応関係を、ネームサーバを通して知るというものである。

【００９３】各パスがどのキャッシュを使用しているかを確認したら、使用しているキャッシュ毎にパスをグループ化する（ステップ３０３）。このグループ化したパス群はパスグループとして扱われることになる。最後に、パスグループの情報をパスグループテーブル１４１に記述する（ステップ３０４）。

【００９４】図４は、本実施の形態のマッピング管理部１５０の動作を説明するためのフローチャートである。マッピング管理部１５０は、まずホストコンピュータのブート時にスクリプト等により立ち上げられる。そして、まずファイルシステム１２０が担当するＩ／Ｏ要求のアドレス情報の範囲を特定する（ステップ４０１）。具体的には、ＬＵＮやＬＢＡが使用されている範囲が何番から何番までかを特定するということである。

【００９５】次に、アドレス情報とパスグループのマッピングを行なう（ステップ４０２）。ステップ４０２のマッピングの目的は、シーケンシャルなリードＩ／Ｏを単一のパスグループに導きつつ、ランダムなＩ／Ｏはなるべくそのような制限を設けずに、従来と同様な負荷分散アルゴリズムを用いてマルチパス負荷分散を行なうことができるようにすることにある。具体的にはＩ／Ｏ要求のアドレス情報であるＬＵＮとＬＢＡから、一意のパスグループが導かれるように行なう。以下に例を２つ示す。

【００９６】１つ目は論理ボリューム番号（以下ＬＵＮ）を用いる方法である。この場合は、各ＬＵＮを適当に（例えば昇順に）パスグループに対してマッピングする。この方法は、シーケンシャルなＩ／Ｏはほとんどどの通信ポートを用いているか特定する。そして、その通信ポートを使っているホスト制御部がどのキャッシュを使っているか調べる。

【００９７】２つ目は論理アドレス（以下ＬＢＡ）の区画を用いる方法である。この場合は、ＬＢＡを或る適当な幅毎に区画化し、各区画を適当に（例えば昇順に）パスグループに対してマッピングしていく。この方法は、シーケンシャルなＩ／ＯはＬＢＡがかなり接近しているという特性を利用している。すなわち、シーケンシャルなＩ／Ｏはほとんど同じパスグループにマッピングされることになる。

【００９８】また、ここではいずれか１つを用いる方法について説明したが、ＬＵＮとＬＢＡを組み合わせてパスグループを導く方法も考えられる。

【００９９】ステップ４０２においてＬＢＡの区画を用いる場合は、その区画幅の決め方として２つ方法が考えられる。つまり、データ記憶装置に教えてもらう方法と、ネームサーバ等の第三者に教えてもらう方法とがある。

【０１００】１つ目の方法は、例えば各パスがデータ記憶装置のどの通信ポートを用いているか特定する際に、ホストコンピュータとデータ記憶装置間で通信が発生するから、その際にデータ記憶装置側からホストコンピュータへ最適な区画幅を渡すというものである。

【０１０１】２つ目は、ホストコンピュータはネームサーバを通して最適な区画幅の情報を知るというものである。

【０１０２】アドレス情報とパスグループのマッピングを行なったら、その情報をマッピングテーブル１５１に記述する（ステップ４０３）。

【０１０３】次に、本実施の形態の効果について説明する。以上説明したように、本実施の形態によれば、パスグループ管理部１４０、マッピング管理部１５０は合わせて、シーケンシャルなリードＩ／Ｏを単一のキャッシュに送付する制限を加えるための情報を提供している。このような制限を加えることで、シーケンシャルなリードＩ／Ｏは、キャッシュを共有しているホスト制御部群（２１０、２１１）によって処理されるようになる。そして、その上で負荷分散制御部１３０が負荷分散を行なっているので、ホスト制御部２１０、２１１がシーケンシャルなリードＩ／Ｏを選り分ける精度が低下する問題、複数のホスト制御部２１０、２１１から同一のアドレス領域へのプリフェッチを多重に行なってしまう問題は発生しない。

【０１０４】次に、本発明の第２の実施の形態について図面を参照して詳細に説明する。図５は、本発明の第２の実施の形態による負荷分散システムの構成を示すブロック図である。

【０１０５】図５を参照すると、本発実施の形態の負荷分散システムの構成の第１の実施の形態との相違点は、ホストコンピュータ１００ａにおいて、第１の実施の形態におけるパスグループ管理部１４０とマッピング管理部１５０を備えない点である。

【０１０６】すなわち、接続されているホスト制御部２１０、２１１が使用するデータ記憶装置２００のキャッシュ毎にパスをグループ化したパスグループの情報と、Ｉ／Ｏ要求のアドレス情報とパスグループのマッピング情報を、管理者が前もってパスグループテーブル１４１ａ、マッピングテーブル１５１ａとしてホストコンピュータ１００ａに与えておく。

【０１０７】負荷分散制御部１３０ａは、これらのテーブルを参照し、負荷分散アルゴリズムに基づき使用するパスを決定する。

【０１０８】以上説明したように本実施の形態によれば、第１の実施の形態の効果に加えて、情報を人間が設定するために非常に細やかな設定が可能となる点や、システムを非常にシンプルに構成することができるという点がある。

【０１０９】次に、本発明の第３の実施の形態について詳細に説明する。図６は、本発明の第３の実施の形態による負荷分散システムの構成を示すブロック図である。

【０１１０】本発明の第３の実施の形態の負荷分散システムの構成は、第１の実施の形態と同様であるが、ホストコンピュータ１００ｂのマッピング管理部１５０ｂにおいて、定期的にマッピングテーブル１５１を更新する機能を備えるようにした点を特徴とする。

【０１１１】図７は、本実施の形態のマッピング管理部１５０ｂの動作を説明するためのフローチャートである。図７を参照すると本実施の形態のマッピング管理部１５０ｂは、まずホストコンピュータのブート時にスクリプト等により立ち上げられる。そして、まずファイルシステム１２０が担当するＩ／Ｏ要求のアドレス情報の範囲を特定する（ステップ７０１）。

【０１１２】次に、管理者に定められた時刻が来るまで待つ（ステップ７０２）。定められた時刻に達すると、アドレス情報とパスグループのマッピングを行なう（ステップ７０３）。アドレス情報とパスグループのマッピングを行なったら、その情報をマッピングテーブル１５１に記述する（ステップ７０４）。この後は、再び管理者に定められた時刻が来るまでは、待機することになる。

【０１１３】第１の実施の形態と異なる点は、マッピング管理部１５０ｂが、定期的にマッピングテーブル１５１を、負荷分散情報テーブル１３１を参照して更新する点である。例えば、本実施の形態のマッピング管理部１５０ｂは、毎日午前１２時にその日一日のパスの利用状況を参照してマッピングテーブルを更新するような形になる。

【０１１４】定期的にマッピングテーブル１５１を、負荷情報を参照して更新することで、パスの負荷状況に沿ってパスグループとＩ／Ｏのアドレス情報をマッピングすることが可能になる。引いては、より適切なマルチパス負荷分散が可能になる。

【０１１５】以上説明したように本実施の形態によれば、第１の実施の形態の効果に加えて、定期的に負荷情報を参照してマッピングテーブル１５１を更新することで、パスの負荷状況に沿ってパスグループとＩ／Ｏのアドレス情報をマッピングすることが可能になる。パスの負荷状況に沿ったマッピングは、より適切なマルチパス負荷分散を実現する。

【０１１６】次に、本発明の第４の実施の形態について詳細に説明する。

【０１１７】図８は、本発明の第４の実施の形態による負荷分散システムの構成を示すブロック図である。図８を参照すると、本実施の形態の負荷分散システムの構成の第１の実施の形態との相違点は、ホストコンピュータ１００ｃにおいて、パスグループ管理部１４０とパスグループテーブル１４１を備えない点であり、また、マッピング管理部１５０ｃの機能を変更することにより、パスグループという概念を本実施の形態においては取り扱わないことを特徴とする。

【０１１８】本実施の形態のマッピング管理部１５０ｃは、Ｉ／Ｏのアドレス情報とパスを直接マッピングする。そして、負荷分散制御部１３０ｃは、そのマッピング情報を用いることで、シーケンシャルなリードＩ／Ｏを単一のパスに導きつつ、ランダムなＩ／Ｏはなるべくそのような制限を設けずにマルチパス負荷分散を行な

う。

【0119】図9は、本発明の第4の実施の形態の負荷分散制御部130cの動作を示すフローチャートである。負荷分散制御部130cは、まずホストコンピュータのブート時にスクリプト等により立ち上げられる。そして、ファイルシステム120よりI／O要求が来るまでは待機している（ステップ901）。

【0120】ファイルシステム120よりI／O要求が来ると、そのI／O要求のアドレス情報とマッピングテーブル151と負荷分散情報テーブル131を参照して使用するパスを選択する（ステップ902）。マッピングテーブル151には、I／O要求のアドレス情報であるLUNやLBAからパスが一意に選択できるようなマッピング情報が記述されている。負荷分散情報テーブル131には、各パスの利用状況と、各パスグループにおいてどの負荷分散アルゴリズムを用いるかということが記述されている。

【0121】使用するパスが決定されると、適切なデバイスドライバへI／O要求を送付する（ステップ903）。最後に、新しくI／O要求を送付したことで変化したパスの利用状況を負荷分散情報テーブル131に記述する（ステップ904）。この後は再びファイルシステム120よりI／O要求が来るまでは待機することになる。

【0122】図10は、本発明の第4の実施の形態のマッピング管理部150cの動作を表すフローチャートである。マッピング管理部150cは、まずホストコンピュータのブート時にスクリプト等により立ち上げられる。

【0123】そして、まずファイルシステム120が担当するI／O要求のアドレス情報の範囲を特定する（ステップ1001）。具体的には、LUNやLBAが使用されている範囲が何番から何番までかを特定するということである。

【0124】次に、アドレス情報とパスのマッピングを行なう（ステップ1002）。ステップ1002のマッピングの目的は、シーケンシャルなリードI／Oを単一のパスに導きつつ、ランダムなI／Oはなるべくそのような制限を設けずに、従来と同様な負荷分散アルゴリズムを用いてマルチパス負荷分散を行なうことができるようにすることにある。具体的には、I／O要求のアドレス情報であるLUNとLBAから、一意のパスが導かれるように行なう。

【0125】アドレス情報とパスのマッピングを行なったら、その情報をマッピングテーブル151に記述する（ステップ1003）。

【0126】以上説明したように本実施の形態によれば、第1の実施の形態の効果に加えて、パスグループという概念を必要としないことから、システムをシンプルに構成できる。

【0127】次に、本発明の第5の実施の形態について詳細に説明する。

【0128】図11は、本発明の第5の実施の形態による負荷分散システムの構成を示すブロック図である。図11を参照すると、本発明の第5の実施の形態は、ホストコンピュータ100と、非共有キャッシュ方式のホスト制御装置400、データ記憶装置200d、ネットワーク300を備える。

【0129】図11を参照すると、本実施の形態の負荷分散システムの構成の第1の実施の形態との相違点は、仮想ボリューム機能やキャッシング機能を行なうホスト制御部410、411を擁するホスト制御装置400を、ディスクアレイ240を備えるデータ記憶装置200dと独立させた点である。本実施の形態において、ホスト制御装置400において、データ記憶装置200dの仮想ボリューム機能やキャッシング機能を提供する。このため、データ記憶装置200d自身には、ホスト制御部を備える必要はない。

【0130】本実施の形態のホストコンピュータ100の処理は、第1の実施の形態と同様である。ただし、本実施の形態においてマッピングを行なう物理的パスは、ホストコンピュータ100とホスト制御装置400との間における物理的パスである。

【0131】つまり、ホストコンピュータ100とホスト制御装置400との間における複数の物理的パスを、ホスト制御装置400内のキャッシュ毎にグループ化してパスグループとして管理し、パスグループの情報を示すパスグループ情報を生成する。そして、入出力要求のアドレス情報とパスグループとのマッピングを行ない、シーケンシャルな読み出し要求を単一のパスグループにマッピングし、マッピングの内容を示すマッピング情報を生成する。そして、生成されたマッピング情報を参照して、第1の実施の形態において説明されたように、入出力要求を複数本の物理的パスに適切に分散する。

【0132】なお、図11の例においては、ホスト制御装置400は簡便のため1つのみ示したが、複数のホスト制御装置が連携して動作している場合も考えられる。

【0133】ホスト制御部410、411は、簡便のため2つのみ示したが、2～20程度備えている場合も考えられる。また、各ホスト制御部410、411は3つずつ通信ポートを持っているが、1～10程度備えている場合も考えられるし、ホスト制御部間410、411で異なる数の通信ポートを持つ場合や、異なる性能の通信ポートが混在する場合も考えられる。また、この図11では、1つのホスト制御部が1つのキャッシュを持つ構成になっているが、複数のホスト制御部が1つのキャッシュを共有する場合も考えられる。

【0134】またホスト制御装置400は、ディスクアレイ、ハードディスク、半導体メモリ等のデータ記憶装置を持つ場合も考えられる。

【0135】またデータ記憶装置200dは、簡便のためディスクアレイのみ示したが、ホスト制御部、キャッシュを持つ場合も考えられる。また、データ記憶装置200dは、簡便のため通信ポートを1つのみ示したが、2～50程度備えている場合も考えられる。

【0136】本実施の形態によれば、第1の実施の形態の効果に加えて、スト制御部を擁するホスト制御装置とディスクアレイを備えるデータ記憶装置とを独立の装置することができ、データ記憶装置の構成には手を加えずに、ホスト制御装置を加えたり削除したりする等の変更が容易になるため、ホスト制御部が実現する仮想ボリューム機能、キャッシング機能の性能をより柔軟に調整することが可能になる。

【0137】また、上記各実施の形態の負荷分散システムは、互いに自由に組み合わせて実施することができる。例えば、第5の実施の形態において、第2、3、4の実施の形態のホストコンピュータを用いて実施すること等が可能である。

【0138】例えば、第5の実施の形態において、第4の実施の形態のホストコンピュータ100cを用いる場合には、入出力要求のアドレス情報と、ホストコンピュータ100cとホスト制御装置400との間における複数の物理的パスとのマッピングを行ない、シーケンシャルな読み出し要求を単一のパスにマッピングし、マッピングの内容を示すマッピング情報を生成する。そして、生成されたマッピング情報を参照して、入出力要求を複数本の物理的パスに適切に分散する。

【0139】なお、上記各実施の形態の負荷分散システムは、ホストコンピュータ100、100a、100b、100cにおける負荷分散制御部130、130a、130c、パスグループ管理部140、マッピング管理部150、150b、150cや、その他の機能をハードウェア的に実現することは勿論として、各機能を備えるコンピュータプログラムである負荷分散プログラム90、90a、90b、90c、90dを、コンピュータ処理装置のメモリにロードされることで実現することができる。この負荷分散プログラム90、90a、90b、90c、90dは、磁気ディスク、半導体メモリその他の記録媒体に格納される。そして、その記録媒体からコンピュータ処理装置にロードされ、コンピュータ処理装置の動作を制御することにより、上述した各機能を実現する。

【0140】以上好ましい実施の形態及び実施例をあげて本発明を説明したが、本発明は必ずしも上記実施の形態及び実施例に限定されるものではなく、その技術的思想の範囲内において様々に変形して実施することができる。

【0141】

【発明の効果】以上説明したように本発明の負荷分散システム、負荷分散システムのホストコンピュータ、及び負荷分散プログラムによれば、以下のような効果が達成される。

【0142】第1に、本発明によれば、各パスが接続されているホスト制御部がデータ記憶装置のどのキャッシュを使用しているかを確認し、使用しているキャッシュ毎にパスをグループ化するパスグループ管理部と、I／O要求（入出力要求）のアドレス情報とパスグループのマッピングを行なうマッピング管理部により、シーケンシャルなリードI／Oを単一のキャッシュに送付する制限を加えた上でマルチパス負荷分散を行なうことができる。すなわち、ホスト制御部がシーケンシャルなリードI／Oを選り分ける精度が低下する問題、複数のホスト制御部から同一のアドレス領域へのプリフェッチを多重に行なってしまう問題は発生しない。

【0143】第2に、本発明の第2の実施の形態によれば、パスグループの情報と、I／O要求のアドレス情報とパスグループのマッピング情報を、管理者が前もってパスグループテーブル、マッピングテーブルとしてホストコンピュータに与えておくことで、システムをシンプルに構成することができる。

【0144】第3に、本発明の第3の実施の形態によれば、定期的に負荷情報を参照してI／O要求のアドレス情報とパスグループのマッピングを更新することで、パスの負荷状況に沿ってパスグループとI／Oのアドレス情報をマッピングすることができる。パスの負荷状況に沿ったマッピングは、より適切なマルチパス負荷分散を実現する。

【0145】第4に、本発明の第4の実施の形態によれば、パスグループという概念を用いずに、I／O要求のアドレス情報とパスを直接マッピングすることで、システムをシンプルに構成することができる。

【図面の簡単な説明】

【図1】 本発明の第1の実施の形態による負荷分散システムの構成を示すブロック図である。

【図2】 本発明の第1の実施の形態の負荷分散制御部の動作を説明するためのフローチャートである。

【図3】 本発明の第1の実施の形態のパスグループ管理部の動作を説明するためのフローチャートである。

【図4】 本発明の第1の実施の形態のマッピング管理部の動作を説明するためのフローチャートである。

【図5】 本発明の第2の実施の形態による負荷分散システムの構成を示すブロック図である。

【図6】 本発明の第3の実施の形態による負荷分散システムの構成を示すブロック図である。

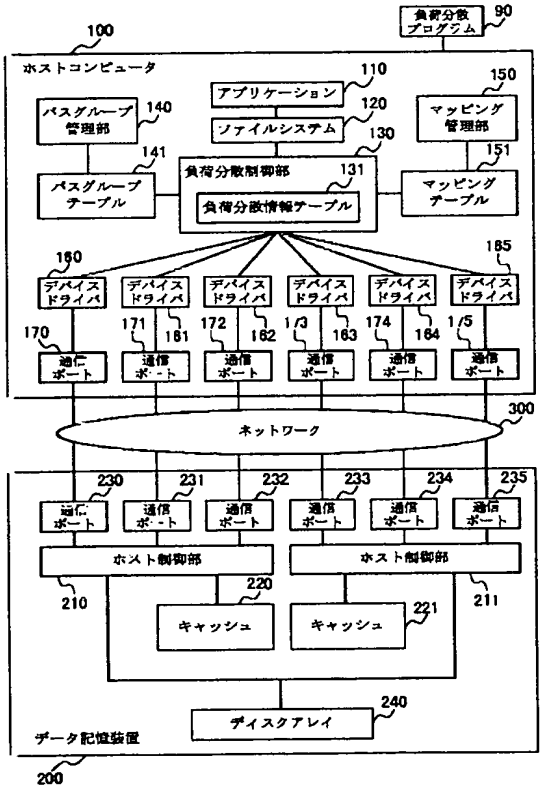
【図7】 本発明の第3の実施の形態のマッピング管理部の動作を説明するためのフローチャートである。

【図8】 本発明の第4の実施の形態による負荷分散システムの構成を示すブロック図である。

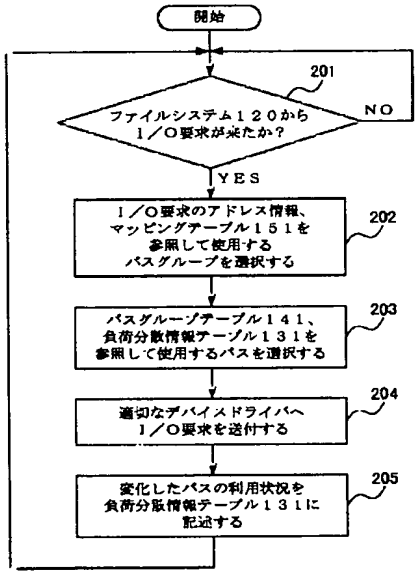
【図9】 本発明の第4の実施の形態の負荷分散制御部の動作を説明するためのフローチャートである。

【図１０】　本発明の第４の実施の形態のマッピング管理部の動作を説明するためのフローチャートである。
【図１１】　本発明の第５の実施の形態による負荷分散システムの構成を示すブロック図である。
【図１２】　従来の負荷分散システムの構成を示すブロック図である。
【図１３】　従来の共有キャッシュ方式と非共有キャッシュ方式のそれぞれのデータ記憶装置の構成を示すブロック図である。
【符号の説明】
９０、９０ａ、９０ｂ、９０ｃ、９０ｄ　負荷分散プログラム
１００、１００ａ、１００ｂ、１００ｃ　ホストコンピュータ
１１０　アプリケーション
１２０　ファイルシステム
１３０、１３０ａ、１３０ｃ　負荷分散制御部
１３１　負荷分散情報テーブル
１４０　パスグループ管理部
１４１、１４１ａ　パスグループテーブル
１５０、１５０ｂ、１５０ｃ　マッピング管理部
１５１、１５１ａ　マッピングテーブル
１６０～１６５　デバイスドライバ
１７０～１７５　通信ポート
２００、２００ｄ　データ記憶装置
２１０、２１１、４１０、４１１　ホスト制御部
２２０、２２１、２２０、２２１　キャッシュ
２３０～２３５、４３０～４３５　通信ポート
２４０　ディスクアレイ
３００　ネットワーク
４００　ホスト制御装置

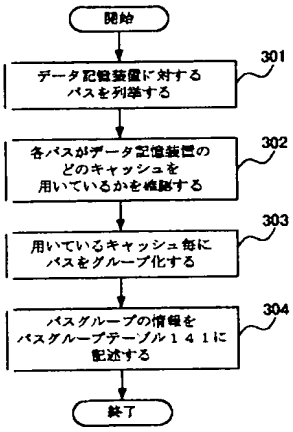
【図１】



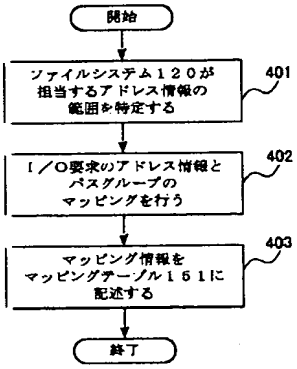
【図２】



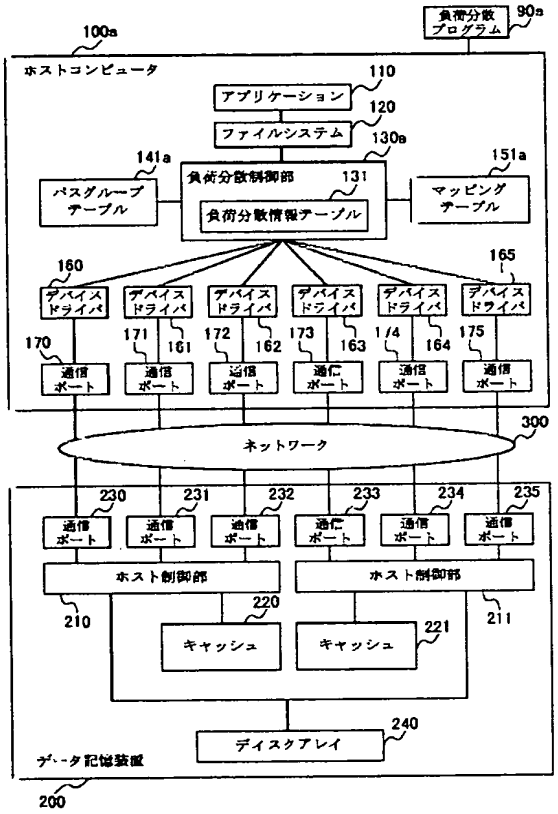
【図3】


開始
データ記憶装置に対する
パスを列挙する
301
各パスがデータ記憶装置の
どのキャッシュを
用いているかを確認する
302
用いているキャッシュ毎に
パスをグループ化する
303
パスグループの情報を
パスグループテーブル１４１に
記述する
304
終了

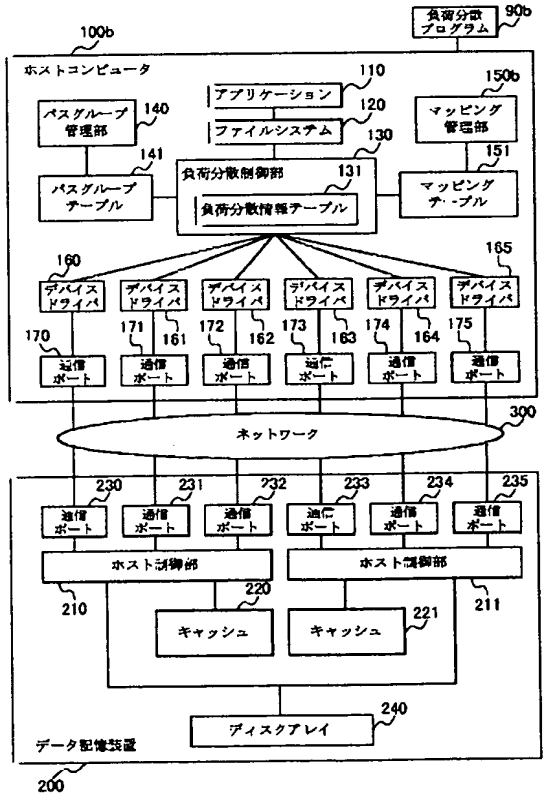
【図4】


開始
ファイルシステム１２０が
担当するアドレス情報の
範囲を特定する
401
Ｉ／Ｏ要求のアドレス情報と
パスグループの
マッピングを行う
402
マッピング情報を
マッピングテーブル１５１に
記述する
403
終了

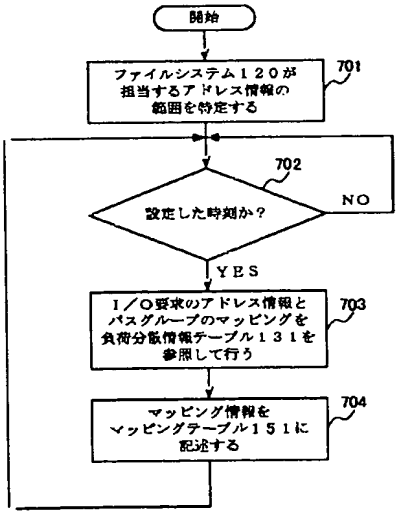
【図5】


負荷分散
プログラム
90a
100a
ホストコンピュータ
110
アプリケーション
120
ファイルシステム
130a
141a
パスグループ
テーブル
負荷分散制御部
131
負荷分散情報テーブル
151a
マッピング
テーブル
160
165
デバイス
ドライバ
170
171
161
172
162
173
163
174
164
175
通信
ポート
300
ネットワーク
230
231
232
233
234
235
ホスト制御部
210
211
220
キャッシュ
221
240
ディスクアレイ
データ記憶装置
200

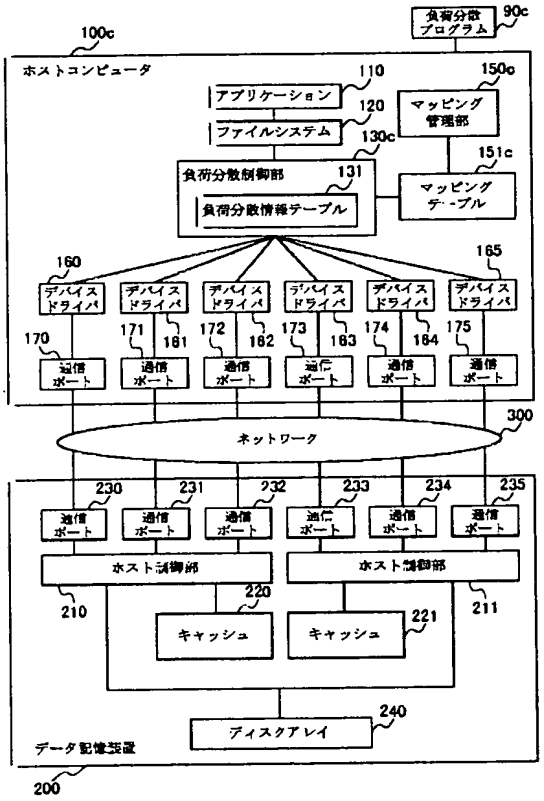
【図6】


負荷分散
プログラム
90b
100b
ホストコンピュータ
110
アプリケーション
120
ファイルシステム
130
140
パスグループ
管理部
150b
マッピング
管理部
141
パスグループ
テーブル
負荷分散制御部
131
負荷分散情報テーブル
151
マッピング
テーブル
160
165
デバイス
ドライバ
170
171
161
172
162
173
163
174
164
175
通信
ポート
300
ネットワーク
230
231
232
233
234
235
ホスト制御部
210
211
220
キャッシュ
221
240
ディスクアレイ
データ記憶装置
200

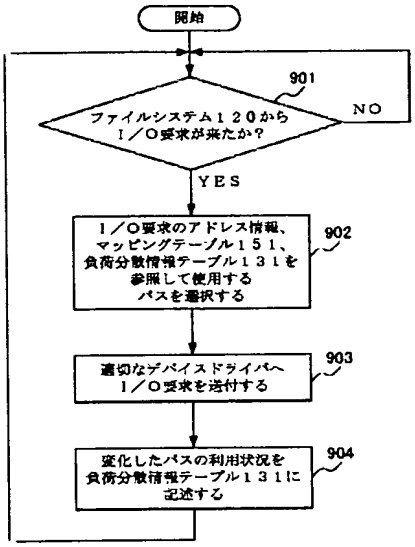
【図7】


開始
ファイルシステム１２０が
担当するアドレス情報の
範囲を特定する
701
702
設定した時刻か？
NO
YES
Ｉ／Ｏ要求のアドレス情報と
パスグループのマッピングを
負荷分散情報テーブル１３１を
参照して行う
703
マッピング情報を
マッピングテーブル１５１に
記述する
704

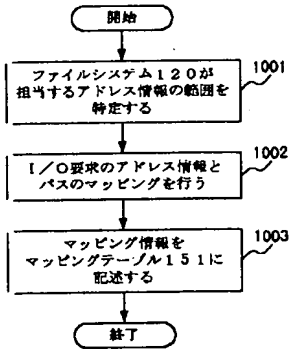
【図8】


負荷分散プログラム
90c
100c
ホストコンピュータ
アプリケーション
110
ファイルシステム
120
マッピング管理部
150c
負荷分散制御部
131
負荷分散情報テーブル
130c
マッピングテーブル
151c
160
165
デバイスドライバ
170
171
161
172
162
173
163
174
164
175
通信ポート
ネットワーク
300
230
231
232
233
234
235
ホスト制御部
210
ホスト制御部
211
220
キャッシュ
221
キャッシュ
240
ディスクアレイ
データ記憶装置
200

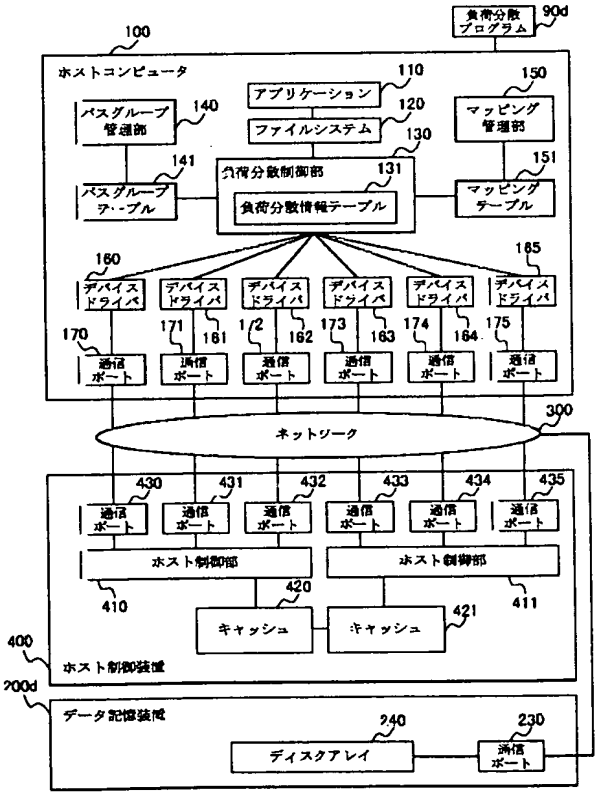
【図9】


開始
901
ファイルシステム120から I/O要求が来たか?
NO
YES
I/O要求のアドレス情報、マッピングテーブル151、負荷分散情報テーブル131を参照して使用するパスを選択する
902
適切なデバイスドライバへI/O要求を送付する
903
変化したパスの利用状況を負荷分散情報テーブル131に記述する
904

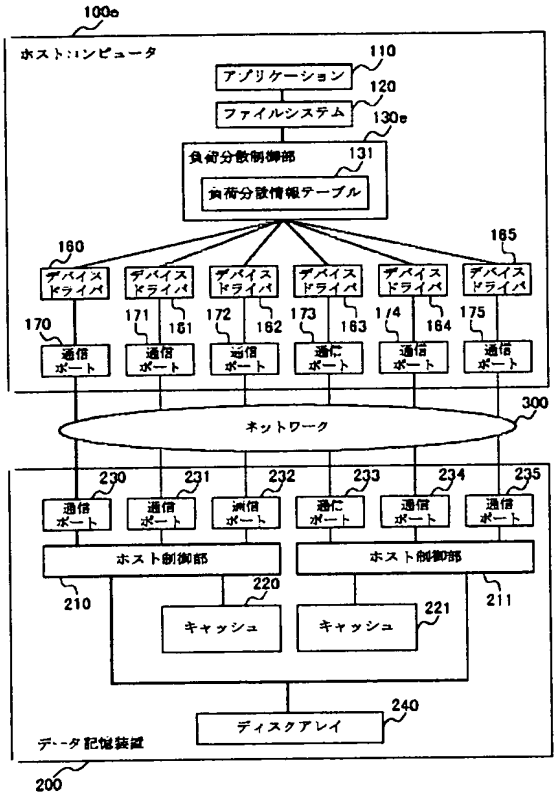
【図10】


開始
ファイルシステム120が担当するアドレス情報の範囲を特定する
1001
I/O要求のアドレス情報とパスのマッピングを行う
1002
マッピング情報をマッピングテーブル151に記述する
1003
終了

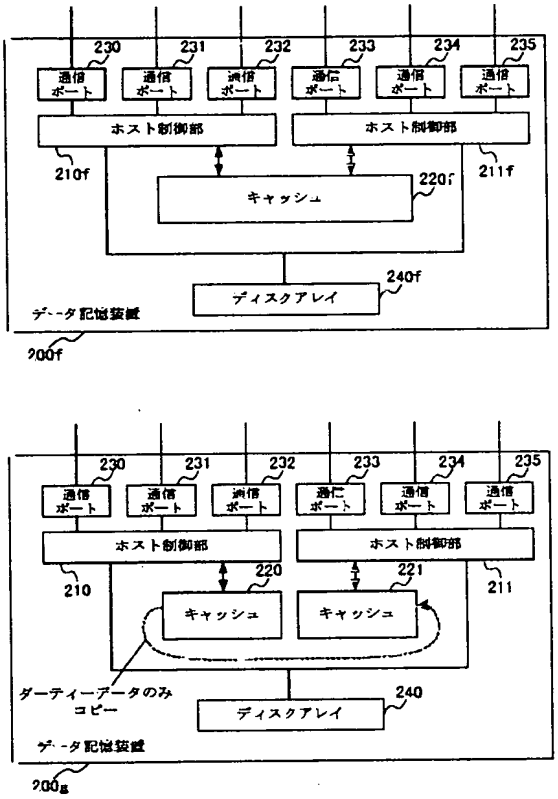
【図11】


負荷分散プログラム
90d
100
ホストコンピュータ
パスグループ管理部
140
アプリケーション
110
ファイルシステム
120
マッピング管理部
150
パスグループテーブル
141
負荷分散制御部
130
負荷分散情報テーブル
131
マッピングテーブル
151
160
165
デバイスドライバ
170
171
161
172
162
173
163
174
164
175
通信ポート
ネットワーク
300
430
431
432
433
434
435
ホスト制御部
410
ホスト制御部
411
420
キャッシュ
421
キャッシュ
400
ホスト制御装置
200d
データ記憶装置
240
ディスクアレイ
230
通信ポート

【図12】



【図13】



フロントページの続き

(51)Int.Cl.[7]	識別記号	FI	(参考)
G06F 12/08	551 557	G06F 12/08	551Z 557